NETFLIX DATA ANALYSIS

Dissertation submitted in fulfilment of the requirements for the Degree of

BATCHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Devarinti Tharun Kumar

Registration number

12212368

K22URA14

Submitted to – Ved Prakash Chaubey

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Sep,2024

Supervisor Certificate

Lovely Professional University

School of Computer Science and Engineering

Certificate of Supervision

This is to certify that the project report titled "Netflix Data Analysis" has been carried out by Devarinti Tharun Kumar (Registration No. 12212368) under my supervision in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab**.**

The project embodies the student's original work and has not been submitted to any other institution or university for the award of any degree or diploma. This report provides an insightful analysis of Netflix's dataset, with particular emphasis on understanding viewing trends, popular genres, content release patterns, and customer preferences, based on exploratory data analysis (EDA) and visualization.

I hereby approve this project report and consider it worthy of submission for evaluation.

Ved Prakash Chaubey

School of Computer Science and Engineering

Lovely Professional University

Signature: _____

Acknowledgement

I would like to extend my profound gratitude towards my University and UpGrad for the brilliant opportunity that was provided to me to work upon this knowledgeable project on "Netflix Data Analysis." This project helped in enhancing my data analysis skills a lot and gave more insight into the trends of Netflix data, thus helping me in my academic as well as professional growth.

The Netflix EDA project has opened my eyes to understanding the streaming industry and its economic and social impact. Data-driven insights have highlighted trends and viewing patterns that could, on some occasions, be valuable for framing future strategies related to content creation, audience engagement, and enhancing the platform's overall user experience.

I would not want to miss this opportunity to extend my deepest appreciation to all the people who supported me in this project. I really appreciate different agencies and open data platforms for the availability of the netflix dataset, which enabled me to do an extensive analysis.

I value very much that such advanced analysis and development of efficient visuals on netflix trends were possible with sophisticated data analysis tools like Pandas, Matplotlib, and Seaborn.

I would like to express my deep gratitude to my mentor, Mr. Ved Prakash Chaubey, for the limitless guidance and feedback given by him with respect to the project. His expertise and insight were really important in refining my analysis and making sure that the findings are correct and profound enough.

First and foremost, I would like to thank my family and friends for their support and encouragement through steady belief in the importance of this project.

Table of Contents

Title: Netflix Data Analysis

Abstract:

The dataset, titled Netflix, serves as the foundation for this analysis, offering detailed insights into Netflix’s content library, including titles, genres, release years, ratings, regional contributions, and durations. The primary objective of this project is to analyse content trends and viewer preferences to uncover patterns that optimize Netflix's content strategy and audience satisfaction. Data cleaning ensured accuracy and reliability, while feature engineering (e.g., "Content Age" and "Genre Popularity Index") and feature scaling enhanced analysis depth. Advanced visualizations using Matplotlib and Seaborn, such as correlation matrices, pair plots, and PCA, revealed significant findings, including a sharp increase in content production from 2015, the dominance of Dramas and Comedies, higher ratings for Documentaries, and the growing contributions of countries like India and South Korea. The analysis also highlighted seasonal release trends favouring the fourth quarter and strong audience preferences for specific genres and regions, providing actionable insights for strategic decisions on content creation, licensing, and audience engagement.

Problem Statement

In the competitive streaming industry, understanding viewer preferences, content trends, and regional dynamics is essential for sustained success. Netflix, one of the largest global streaming platforms, provides a wealth of data on its content library, including genres, release years, audience ratings, and regional contributions. By analyzing this data, strategic insights can be gained into content popularity, viewer engagement, and regional contributions to its diverse catalogue.

The primary goal of this project is to explore and analyze Netflix's content dataset to:

- Identify trends in genre popularity, release year distributions, and regional contributions.
- Uncover patterns in audience ratings and their correlation with genres, content age, and regions.
- Analyze the relationships between content features to provide actionable insights for optimizing content strategy and enhancing audience engagement.

This analysis will help Netflix better understand its audience, refine its content offerings, and develop effective strategies for content production, licensing, and targeted marketing in a rapidly evolving global streaming landscape.

Data Description

The dataset used for this analysis is a comprehensive collection of Netflix content information, providing insights into various attributes of titles available on the platform. Each entry includes key features that help understand the content's popularity, ratings, and production trends. The main features in the dataset include:

- **Title**: Tells us the name of the movie / series
- **Genre**: Describes the genre of the movie / series
- **Tag**: is a list of tags on the movie / series
- **Language**: Indicates in which language the film / series is available
- **Series or Movie**: Indicates whether this product is a TV series or a movie
- **Hidden Gem Score**: This is a score that suggests if this movie / series is a hidden gem (i.e. something that is not mainstream but is a great product and may deserve more attention)
- **Country Availability**: Please indicate in which country this film / series is available
- **Runtime**: Indicates the duration of the movie or an episode of the series
- **Director**: Tells us how is the director of this product
- **Writer**: Tells us how is the writer of this product,
- **Actors**: Tells us how are the main actors of this product,
- **View Rating**: Indicates the rating of the product,
- **IMDb Score**: This is the score given to the product by IMDb,
- **Rotten Tomatoes Score**: This is the score given to the product by Rotten Tomatoes,
- **Awards Received**: It tells us how much awards this film/serie has received ,

- **Awards Nominated For**: It tells us for which awards this film/movie has been nominated,
- **Release Date**: It indicates when this product has been released,
- **Netflix Release Date**: It indicates when this product has been released on Netflix,
- **Summary**: It tells us a little summary of the plot of the product,
- **IMDb Votes**: It indicates the scores given by IMDb,
- **Image**: An Image
- **Released_Year**: It indicates the year the product was released
- **Released_Year_Net**: It indicates the year the product was released on Netflix

Data Source

The dataset was sourced from **Kaggle** and is intended for academic and analytical purposes. It provides a rich foundation for exploring trends in Netflix's content library, analyzing viewer preferences, and identifying patterns in content production, genres, and ratings.

Data Preprocessing Steps

The raw dataset contains some inconsistencies and missing values, particularly in pricing and ratings**.**

Preprocessing Steps Included

1. **Handling Missing Values:** Missing values in key columns like Genre, Country, and Rating were addressed. For categorical columns, missing entries were filled with "Unknown," while numeric columns were imputed using the median to maintain data integrity.
2. **Data Type Conversion**: Converted text-based numeric data such as **Duration** to numeric types. Text columns like **Genres** were standardized for uniform formatting.
3. **Removing Duplicates:** Identified and removed duplicate records to ensure a clean dataset.
4. **Outlier Detection and Handling**: Applied IQR (Interquartile Range) methods to identify and treat outliers in **Ratings** and **Duration** columns.
5. **Data Standardization:** Standardized key attributes such as release year and duration to enable consistent analysis across the dataset.

Solution Approach

1. Data Collection and Inspection

- **Load the Dataset**: Imported the Netflix dataset into the environment using Python's Pandas library.
- **Initial Inspection**: Examined the dataset's shape, column names, and data types using .head(), .info(), and .describe() to understand its structure.
- **Missing Value Analysis**: Checked for missing values in all columns and assessed their distribution for appropriate handling.

2. Data Cleaning and Preprocessing

- **Handling Missing Values**: Filled missing values in essential columns like **Ratings** and **Country** with appropriate imputation techniques.
- **Data Type Conversion**: Converted non-numeric data types (e.g., durations stored as text) into numeric formats for analysis.
- **Outlier Handling**: Used box plots and IQR filtering to identify and address extreme values in numeric fields.
- **Data Standardization**: Removed unnecessary columns and ensured uniform formatting of categorical data.

3. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**: Calculated key statistics (mean, median, standard deviation) for numerical columns such as **Ratings** and **Release Year** to understand data distribution.
- **Visual Analysis**:
    - **Genre Analysis**: Bar charts to identify the most and least common genres.
    - **Rating Trends**: Scatter plots and histograms to explore rating distributions.
    - **Regional Contributions**: Heatmaps to examine country-wise contributions to Netflix's content.

4. Feature Engineering

- **New Feature Creation**:
    - **Content Age**: Derived as 2024 - Release Year.
    - **Genre Popularity Index**: Aggregated average ratings per genre.
- **Normalization and Scaling**: Applied normalization techniques to standardize features like **Duration** for consistent visualizations.

5. Statistical Analysis

- **Correlation Analysis**: Generated a correlation matrix to explore relationships between features like **Ratings**, **Content Age**, and **Duration**.
- **Hypothesis Testing**:
    - **ANOVA**: Compared average ratings across genres to identify significant differences.
    - **Correlation Testing**: Pearson correlation tests to assess relationships between variables like ratings and content duration.

6. Visualization and Interpretation

- **Graphical Insights**:
    - Heatmaps, scatter plots, and box plots were used to interpret patterns and trends.
    - Genre-specific and country-wise visualizations highlighted the distribution and trends of content attributes.
- **Interactive Dashboards** (Optional): Created using Plotly to dynamically explore Netflix data trends.

7. Result Analysis and Business Insights

- **Insight Extraction**: Derived actionable insights on genre popularity, audience preferences, and regional trends.
- **Recommendation Development**: Proposed strategies for content curation and audience engagement based on EDA findings.

8. Conclusion and Future Work

- **Summarize Findings**: Highlighted key insights on content trends, rating distributions, and regional contributions.
- **Future Research**: Suggested integrating textual sentiment analysis from reviews and examining more recent datasets to refine insights

Required Libraries or used libraries

1. Data Manipulation and Analysis

- **Pandas**: Used for data manipulation, cleaning, and organizing structured data into dataframes. Essential functions include reading the CSV file, handling missing values, filtering, and aggregating data.

Code: import pandas as pd

- **NumPy**: Utilized for numerical operations and handling arrays. It's often used to calculate statistics like mean, median, and for handling data transformations.

Code : import numpy as np

2. Statistical Analysis

- **SciPy**: Offers statistical functions and tests, such as correlation tests and ANOVA, to understand relationships and differences between groups.

Code: from scipy import stats

3. Data Visualization

- **Matplotlib**: A foundational plotting library for creating static, publication-quality visualizations. It's useful for line plots, histograms, and scatter plots.

Code : import matplotlib.pyplot as plt

- **Seaborn**: Built on top of Matplotlib, it provides more complex visualizations and easier syntax for statistical plots like box plots, pair plots, and heatmaps.

Code: import seaborn as sns

- **Plotly** (Optional): A library for interactive visualizations, suitable for creating dynamic dashboards and visuals that allow for interactive exploration.

Introduction:

The streaming industry has witnessed remarkable growth over the past decade, driven by the demand for on-demand entertainment, diverse content offerings, and personalized user experiences. Netflix, a global leader in this domain, hosts a vast library of titles spanning multiple genres and regions, generating extensive data that reflects audience preferences, content trends, and regional dynamics. This project aims to analyze Netflix’s content dataset to uncover patterns in genre popularity, release years, regional contributions, and audience ratings. By leveraging data cleaning, Exploratory Data Analysis (EDA), and advanced visualization techniques, the study explores correlations between release years and ratings, the impact of regional content on global appeal, and viewer engagement trends. The analysis provides actionable insights to optimize Netflix's content strategies, enhance user satisfaction, and guide decisions in content production, licensing, and marketing. These findings have broad applications, from tailoring regional content strategies to improving marketing and audience engagement. Future work could integrate user reviews and viewership data for a more comprehensive understanding of audience sentiment and content performance, ensuring Netflix’s competitive edge in the evolving streaming landscape.

Objective:

The objectives of this project are:

Data Exploration and Cleaning

- Examine the Netflix dataset to identify missing, inconsistent, or erroneous values.
- Clean the dataset by handling missing entries, standardizing formats for genres, ratings, and durations, and ensuring data quality for meaningful analysis.

Exploratory Data Analysis (EDA)

- Analyze the distribution of content across genres, release years, and regions.
- Investigate the relationship between audience ratings and content attributes like genres and release years to uncover patterns in viewer preferences.
- Visualize key insights using graphs and charts to explore trends in content production, ratings, and regional contributions.

Genre and Regional Analysis

- Compare the popularity of genres across different regions to assess their global and local appeal.
- Identify regional content contributions and their impact on Netflix's diverse catalog.

Audience Rating Insights

- Study the distribution of audience ratings across various genres and release years to understand patterns in viewer satisfaction.

- Explore the relationship between ratings and the duration of content to identify whether longer or shorter content receives better reviews.

Content Strategy and Market Insights

- Provide actionable insights to optimize Netflix's content production and licensing strategies based on genre and regional trends.
- Identify potential areas for content improvement based on audience feedback and rating trends.
- Suggest data-driven strategies for enhancing user engagement and maintaining competitiveness in the streaming industry.

Future Research and Applications

- Highlight opportunities for further analysis, such as applying sentiment analysis to user reviews for deeper insights into viewer satisfaction.
- Propose enhancements to data collection and analysis methods to provide a more comprehensive understanding of audience behavior and content performance.

Methodology

The methodology for this project follows a structured workflow, starting from data collection and cleaning to exploratory data analysis (EDA) and visualization. The key steps involved are:

1. Data Collection

- **Data Inspection**:
    - Reviewed the dataset's structure, including shape, data types, and content, to gain a basic understanding.
    - Checked for missing values, outliers, and any potential data quality issues using functions like .info(), .describe(), and .isnull().sum().

2. Data Cleaning

- **Handling Missing Values**:
    - Columns with over 70% missing data were considered for removal if necessary.
    - Null values in critical columns like **Genres**, **Ratings**, and **Country** were imputed using the mode for categorical data and the median for numerical data.
- **Data Type Conversion**:
    - Converted text-based numerical data (e.g., duration stored as strings) into appropriate numeric formats for analysis.
- **Duplicate Removal**:
    - Identified and removed duplicate records using the drop_duplicates() function to ensure accuracy.
- **Outlier Detection**:

- Used box plots and interquartile range (IQR) methods to identify and handle outliers in fields like **Ratings** and **Duration**.

3. Exploratory Data Analysis (EDA)

- **Descriptive Statistics**:
  - Generated summary statistics for key features like **Release Year**, **Ratings**, and **Duration** using .describe() to understand central tendencies and variability.
- **Data Visualization**:
  - Created visualizations using Matplotlib, Seaborn, and Plotly to explore trends, such as content distribution by year, genre, and region.
  - Analyzed patterns using bar charts, scatter plots, and histograms.
- **Correlation Analysis**:
  - Computed a correlation matrix to examine relationships between numerical features, such as **Ratings** and **Duration**.

4. Tools and Libraries

- **Pandas**: Used for data manipulation and preprocessing.
- **Matplotlib and Seaborn**: Utilized for creating static visualizations such as bar charts, histograms, and box plots.
- **Plotly**: Employed for interactive visualizations, enabling dynamic exploration of data trends.
- **Python**: The primary programming language for coding and executing the analysis.

5. Statistical Analysis

- **Descriptive Statistics**:

- Summarized key metrics for numerical columns, including mean, median, and standard deviation.

- **Correlation Analysis**:
  - Identified relationships between variables like **Release Year**, **Ratings**, and **Duration** using correlation coefficients and heatmaps.
- **Outlier Detection**:
  - Used visualization techniques like box plots to detect and address extreme values.

6. Graphical Visualizations for Netflix Dataset

1. **Content Distribution by Year and Genre**:
   - Bar charts were created to illustrate the number of titles produced each year and their genre classifications.
2. **Regional Contribution**:
   - Pie charts and heatmaps were used to analyze content contributions by region or country.
3. **Rating Distribution**:
   - Histograms highlighted the spread of ratings, showing whether most content was well-rated or if there were significant outliers.
4. **Correlation Between Ratings and Duration**:
   - Scatter plots revealed trends between content duration and ratings to determine if longer or shorter content performed better.
5. **Top Rated Content**:
   - Bar charts displayed the highest-rated titles, providing insights into popular genres and production teams.
6. **Ratings vs. Release Year**:
   - Line plots analyzed how audience ratings evolved over time.

This structured methodology ensured a comprehensive analysis of the Netflix dataset, enabling actionable insights into content trends, regional contributions, and viewer preferences.

```
import numpy as np
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go

netflix = pd.read_csv("netflix_titles.csv")
netflix.head(10)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Children & Family Movies	Equestria's divided. But a bright-eyed hero be...
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...

```
netflix.tail(10)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	2016	TV-Y7	3 Seasons	Kids' TV	Teen surfer Zak Storm is mysteriously transpor...
8798	s8799	Movie	Zed Plus	Chandra Prakash Dwivedi	Adil Hussain, Mona Singh, K.K. Raina, Sanjay M...	India	December 31, 2019	2014	TV-MA	131 min	Comedies, Dramas, International Movies	A philandering small-town mechanic's political...
8799	s8800	Movie	Zenda	Avadhoot Gupte	Santosh Juvekar, Siddharth Chandekar, Sachit P...	India	February 15, 2018	2009	TV-14	120 min	Dramas, International Movies	A change in the leadership of a political part...
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	December 15, 2016	2012	TV-PG	1 Season	International TV Shows, Romantic TV Shows, TV ...	Strong-willed, middle-class Kashaf and carefre...
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	2015	TV-MA	96 min	Dramas, International Movies, Thrillers	Recovering alcoholic Talal wakes up inside a s...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

netflix.dtypes

```
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object
```

# To removethe Duplicates rows permanently
netflix.drop_duplicates()

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

netflix.isnull()

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	False	False	False	False	True	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	True	False	False	False	False	False	False
3	False	False	False	True	True	True	False	False	False	False	False	False
4	False	False	False	True	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	False	False	False	False	False	False	False	False	False	False	False	False
8803	False	False	False	True	True	True	False	False	False	False	False	False
8804	False	False	False	False	False	False	False	False	False	False	False	False
8805	False	False	False	False	False	False	False	False	False	False	False	False
8806	False	False	False	False	False	False	False	False	False	False	False	False

8807 rows × 12 columns

# To use heatmap to show null values count
sns.heatmap(netflix.isnull())

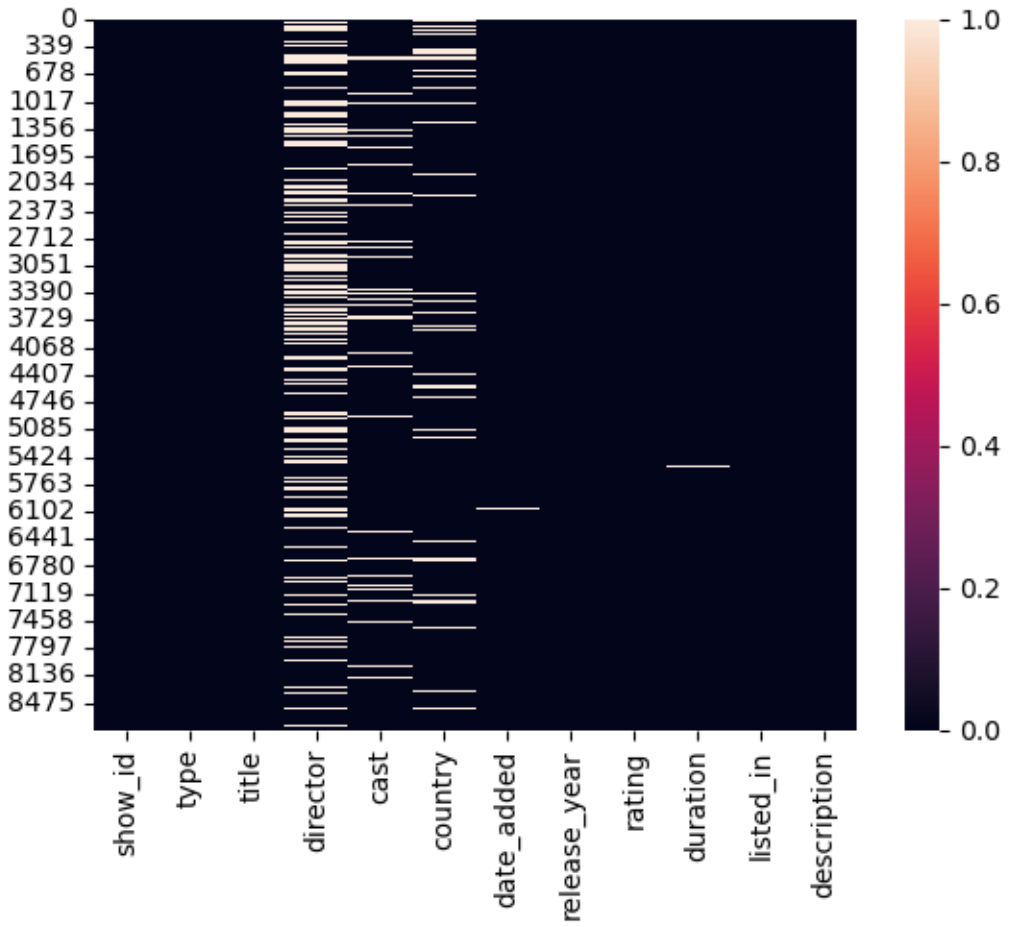

Fig: 1 The image shows a heatmap displaying missing values in a dataset. It highlights the presence of missing data across various columns, with darker colors indicating higher proportions of missing values.

```
# Distribution of content type (Movie vs TV Show)
sns.countplot(data=netflix, x='type')
plt.title('Distribution of Content Type')
plt.show()

# Distribution of release years
plt.figure(figsize=(10, 6))
sns.histplot(netflix['release_year'], bins=20, kde=False)
plt.title('Distribution of Release Years')
plt.show()
```

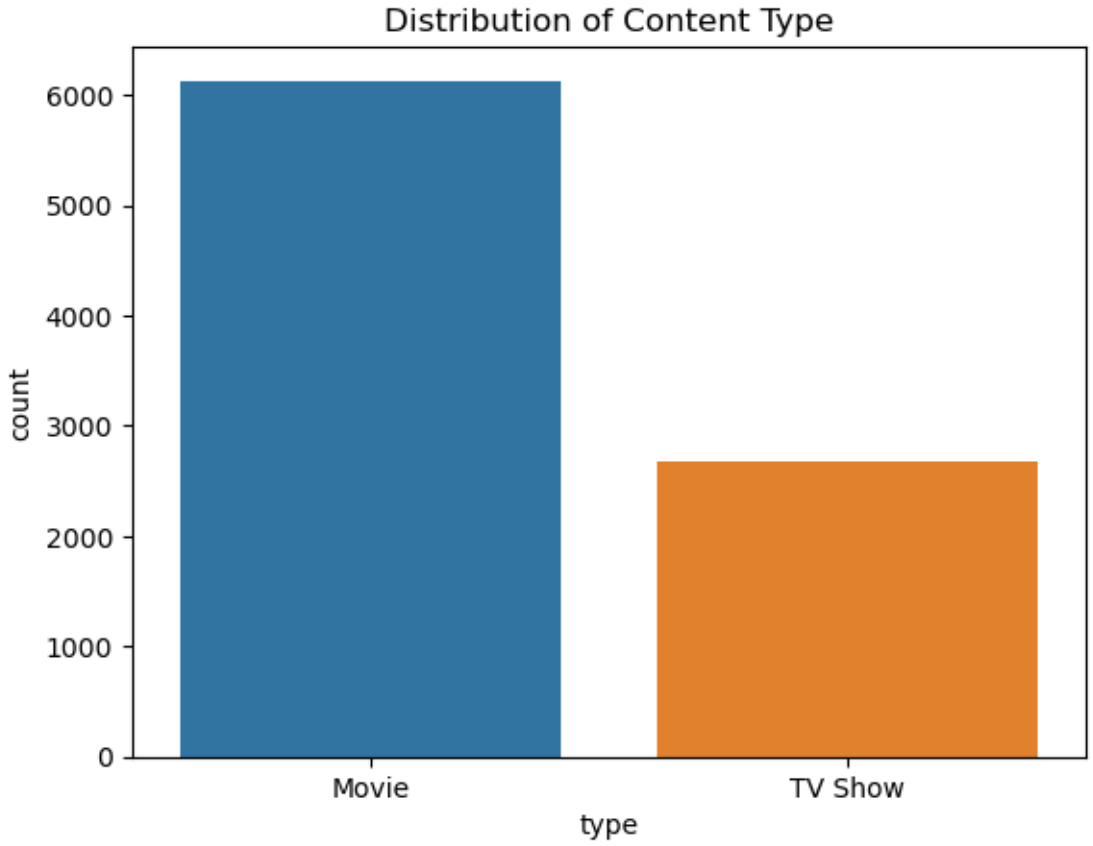

Fig 2 : The bar chart displays the distribution of content types in the dataset, showing a significantly higher count of movies compared to TV shows.

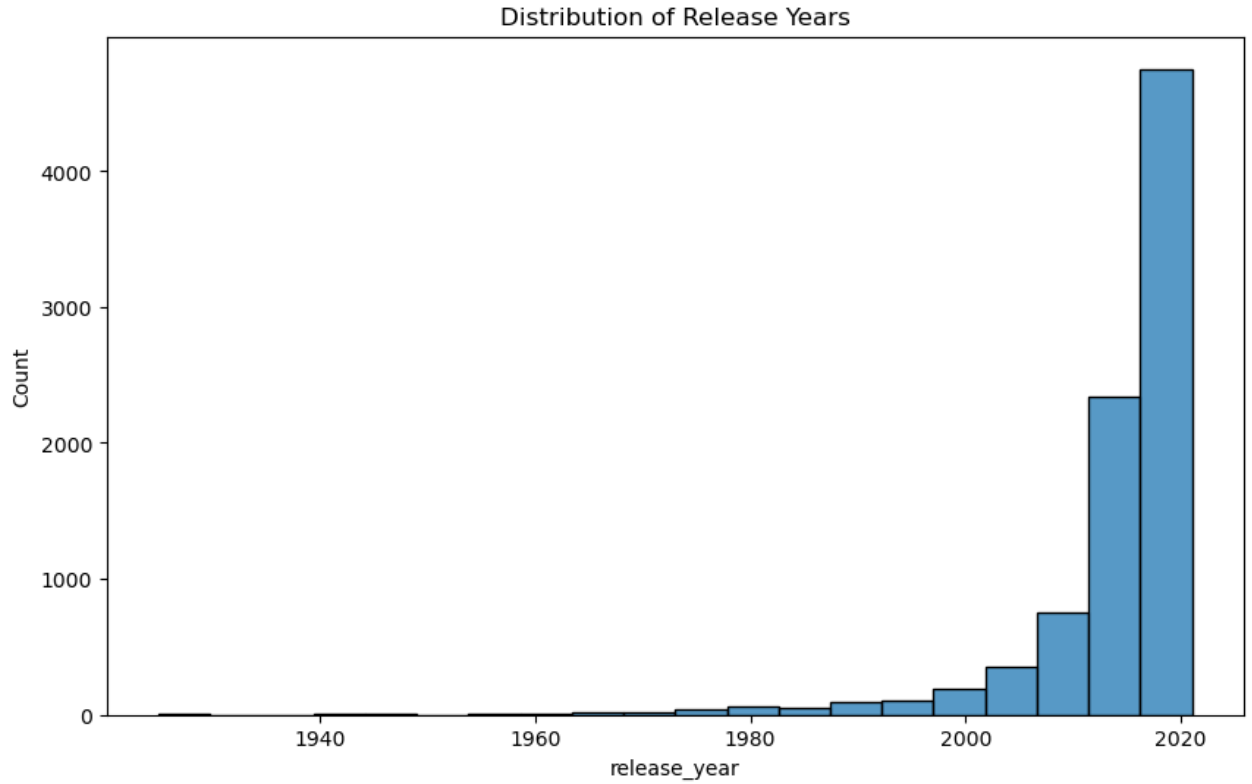

Fig 3 :
The histogram shows the distribution of release years for a dataset, with a clear right skew indicating that most entries are from recent years.

Rating of shows and movies
In [29]:
 # Plot the count of each rating
plt.figure(figsize=(13, 13)) *# Set the figure size*
ax = sns.countplot(x='rating', data=netflix)

 # Rotate x-axis labels for better readability
ax.set_xticklabels(ax.get_xticklabels(), rotation=90, ha="right")

 # Set the title
plt.title('Distribution of Ratings')

 # Show the plot
plt.show()

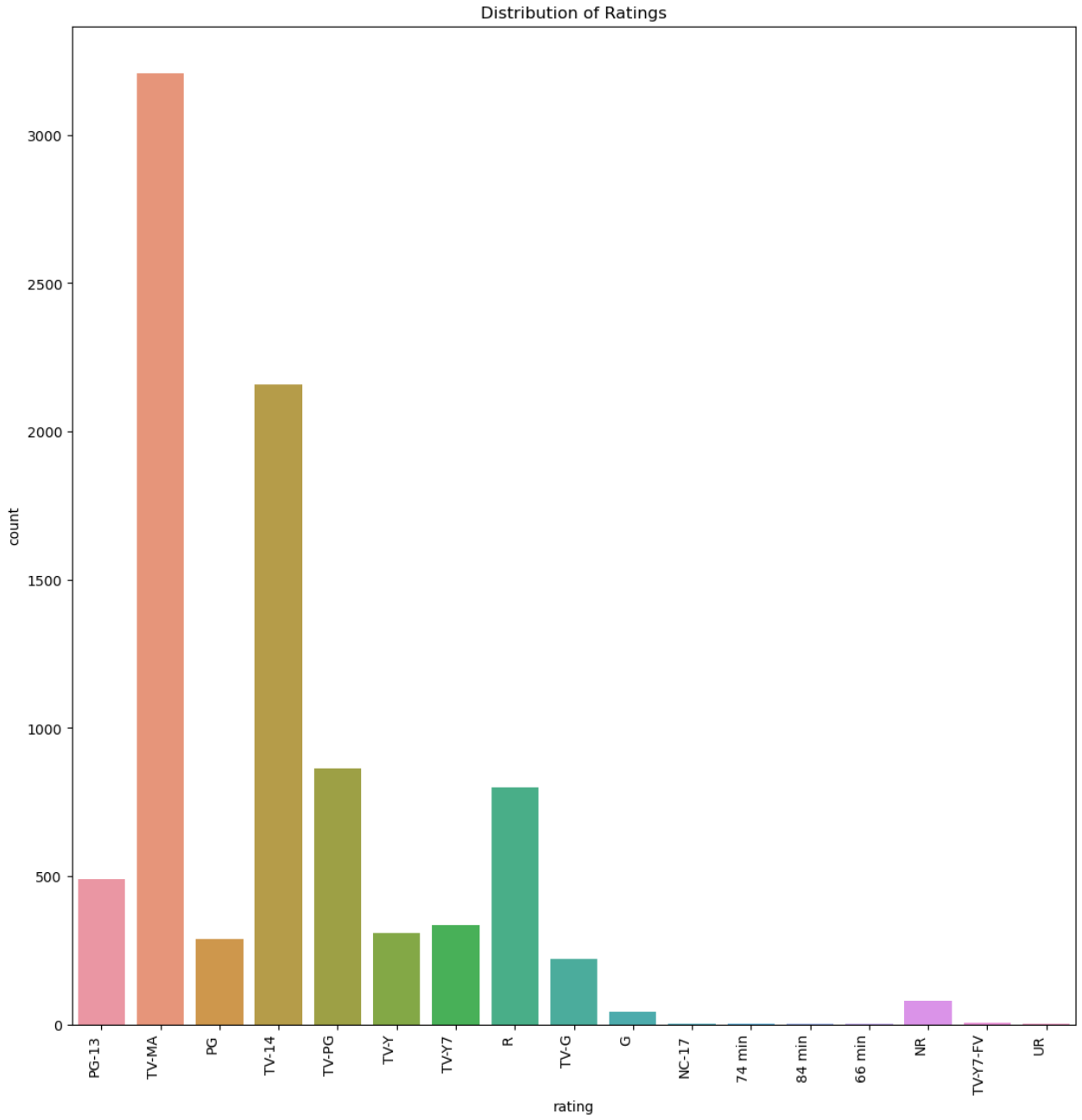

Fig 4 : The plot shows the distribution of ratings for a dataset. The most common ratings are PG-13 and TV-MA.

Relation between Type and Rating

In [30]:

```
plt.figure(figsize=(10,8))
sns.countplot(x='rating',hue='type',data=netflix)
plt.title('Relation between Type and Rating')
plt.show()
```

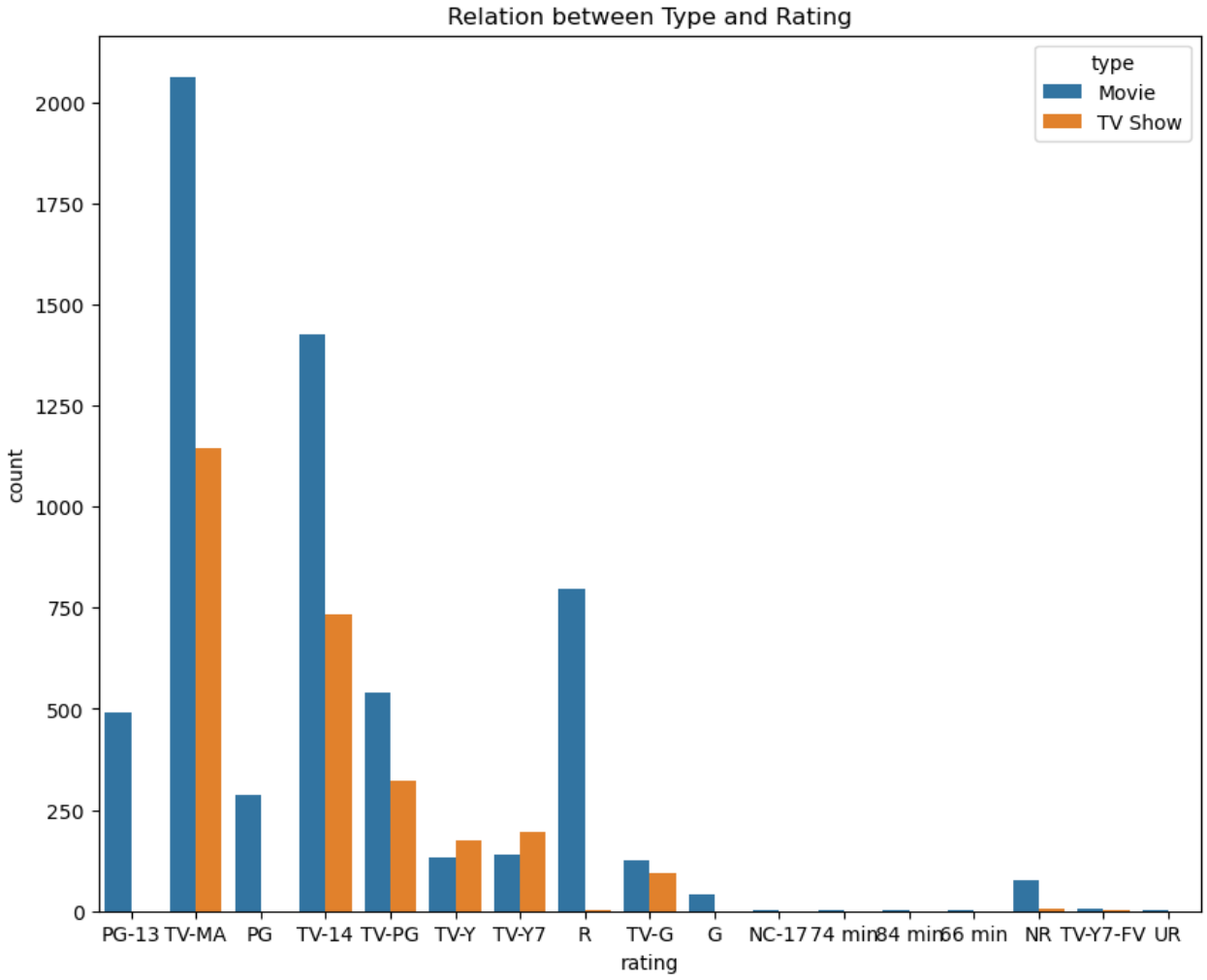

Fig 4 : The plot shows the distribution of ratings for movies and TV shows. The most common rating for movies is PG-13, while the most common rating for TV shows is TV-MA

Bivariate Analysis

In [31]:

```
# Count of movies and TV shows released by year
plt.figure(figsize=(12, 8))
sns.countplot(data=netflix, x='release_year', hue='type')
plt.xticks(rotation=90)
plt.title('Count of Movies and TV Shows Released by Year')
plt.show()
```

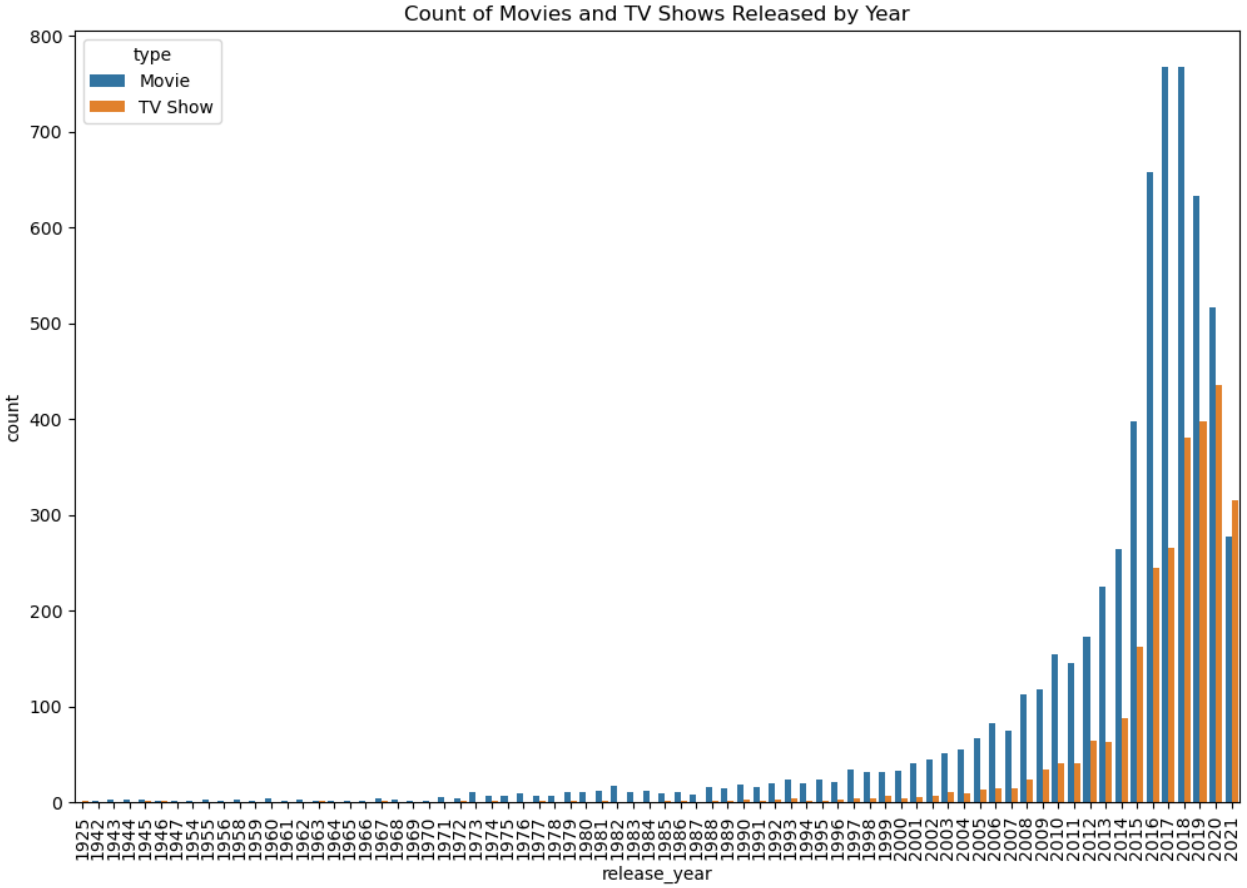

Fig 5: The plot shows the number of movies and TV shows released each year from 1985 to 2021. The number of releases has increased significantly over time, with a particularly sharp increase in recent years.

Line plot for Netflix releases per year

```
release_count = netflix['release_year'].value_counts().sort_index()
plt.figure(figsize=(10, 6))
plt.plot(release_count.index, release_count.values, marker='o')
plt.xlabel('Year')
plt.ylabel('Number of Releases')
plt.title('Netflix Releases Per Year')
plt.show()
```

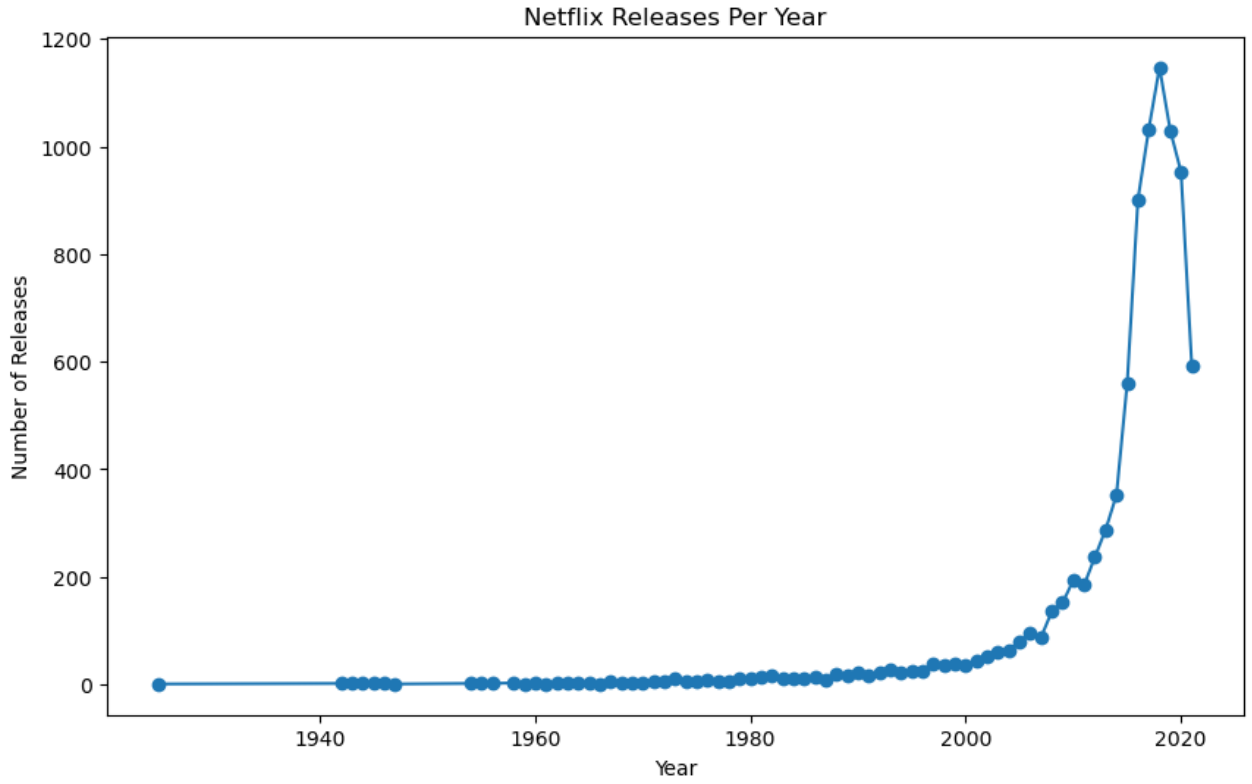

Fig 6 : The line plot shows the number of Netflix releases per year from 1925 to 2021. The number of releases was very low until the late 1990s, after which it started to increase rapidly. In 2021, there were over 1,100 Netflix releases.

```
# Heatmap for country content distribution
top_countries = netflix['country'].value_counts().head(10)
sns.heatmap(top_countries.to_frame(), annot=True, cmap='Blues')
plt.title('Top 10 Countries by Netflix Content Production')
plt.show()
```

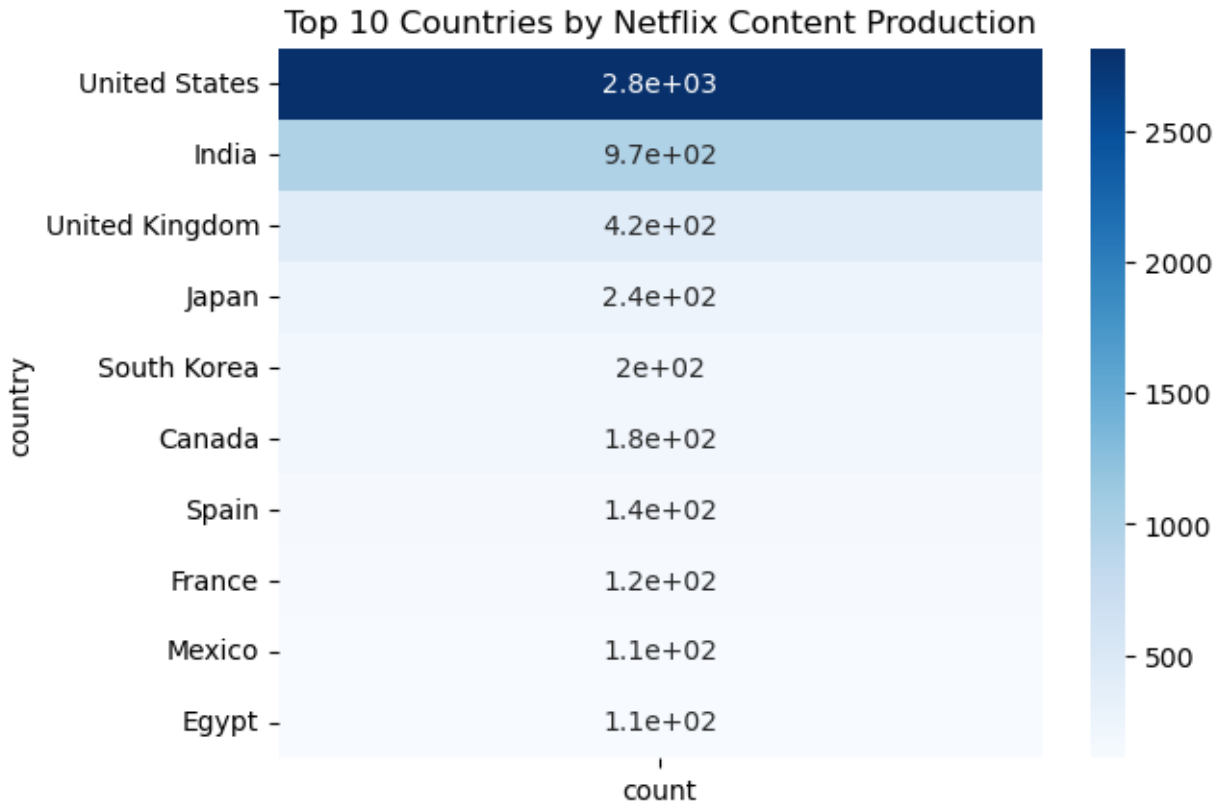

Fig 7 : The bar plot shows the top 10 countries by Netflix content production. The United States has the highest production with over 2800 titles, followed by India with 970 titles.

```
# Bar plot for top genres
top_genres = netflix['listed_in'].value_counts().head(10)
plt.figure(figsize=(10, 6))
top_genres.plot(kind='bar', color='skyblue')
plt.title('Top 10 Genres on Netflix')
plt.show()
```

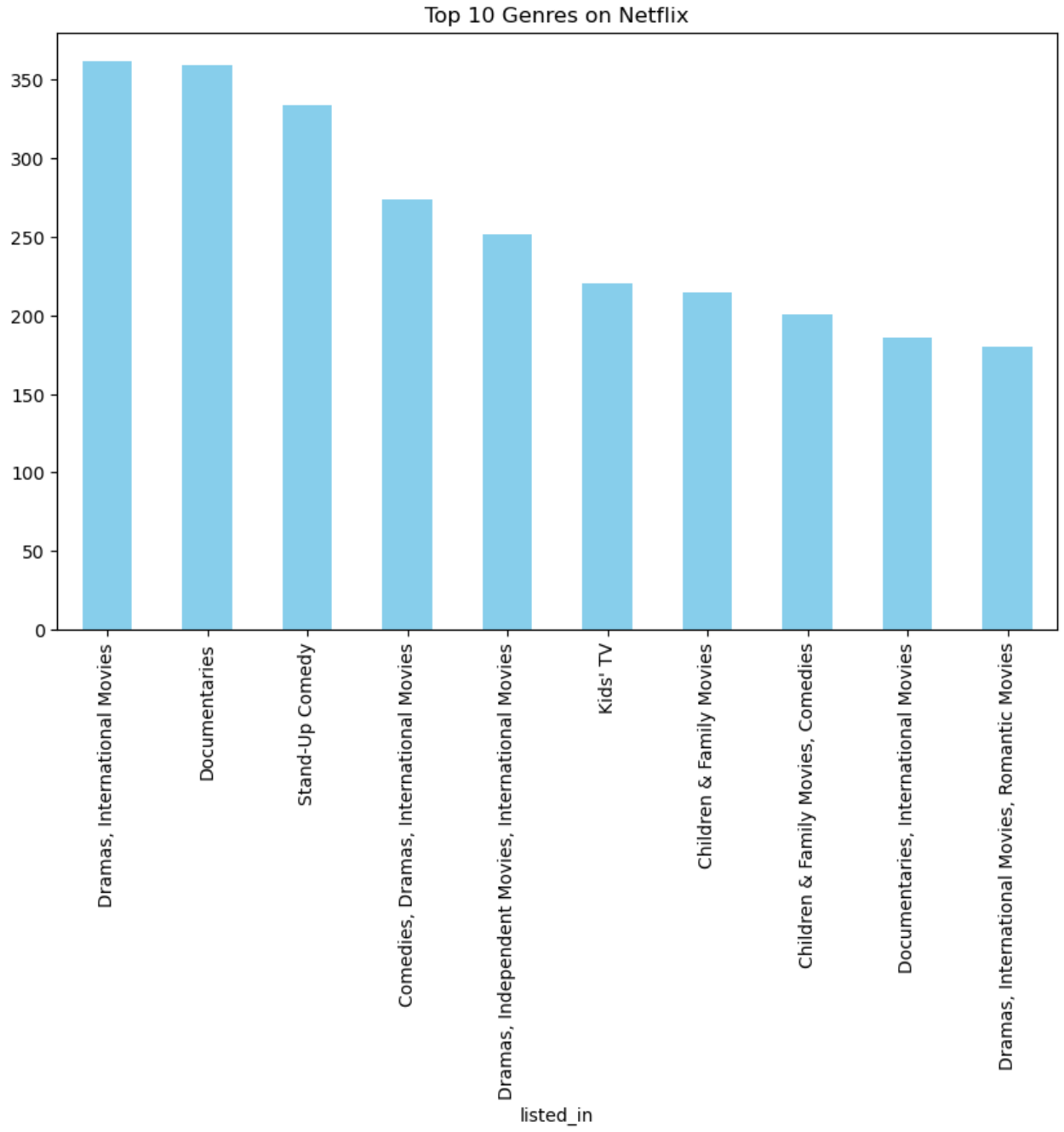

Fig 8 : The bar plot shows the top 10 genres on Netflix. Dramas, International Movies is the most popular genre, with over 350 titles.

Pie-chart for the Type: Movie and TV Shows

In [35]:

```
labels = ['Movie', 'TV show']
size = netflix['type'].value_counts()
colors = plt.cm.Wistia(np.linspace(0, 1, 2))
explode = [0, 0.1]
plt.rcParams['figure.figsize'] = (9, 9)
plt.pie(size,labels=labels, colors = colors, explode = explode, shadow = True, startangle = 90)
plt.title('Distribution of Type', fontsize = 25)
```

```
plt.legend()
plt.show()
```

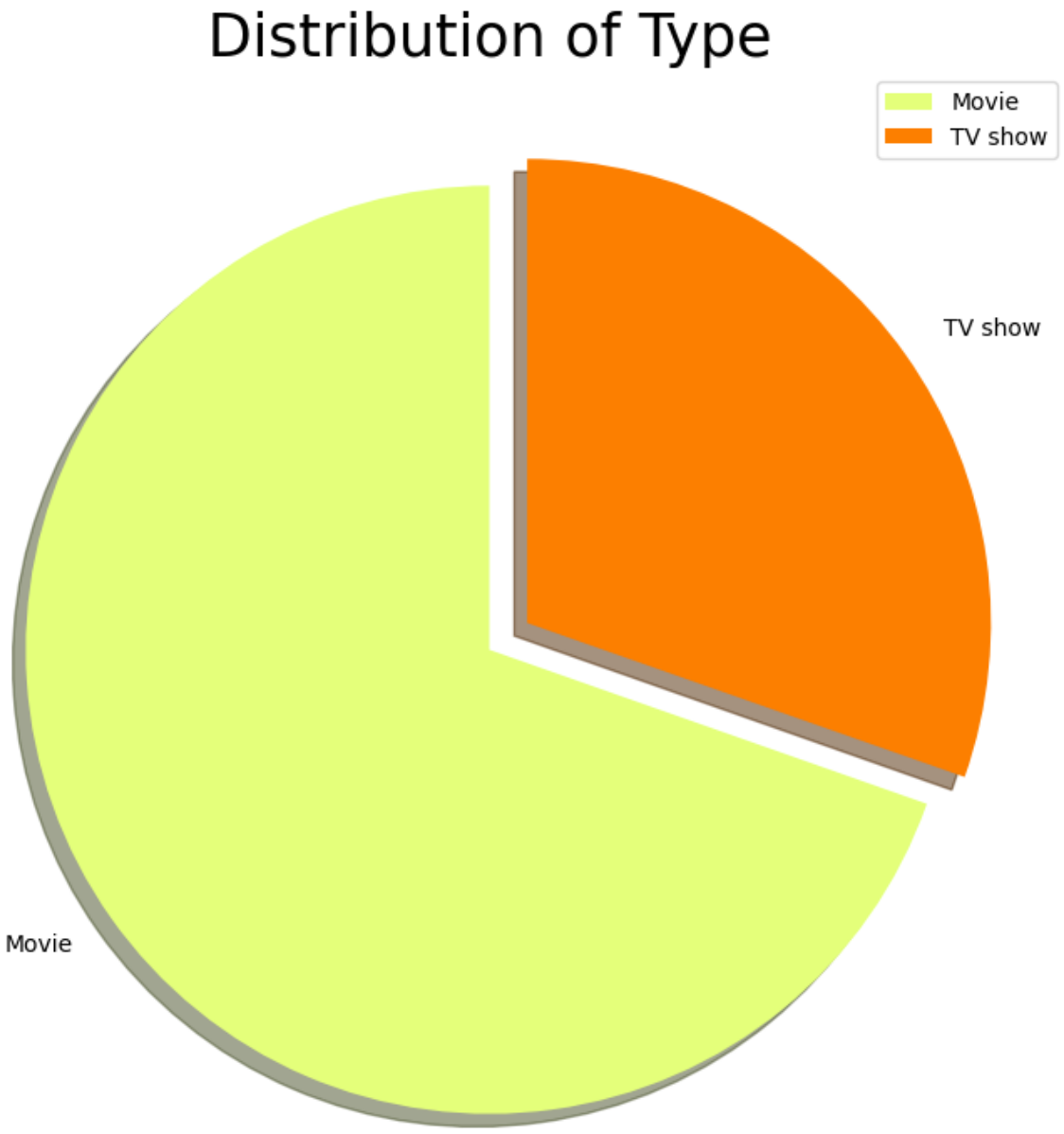

Fig 9 : The pie chart shows the distribution of movie and TV show content on Netflix. Approximately 60% of the content is movies, while 40% is TV shows.

Pie-chart for Rating

In [36]:

```
netflix['rating'].value_counts().plot.pie(autopct='%1.1f%%',shadow=True,figsize=(10,8))
plt.show()
```

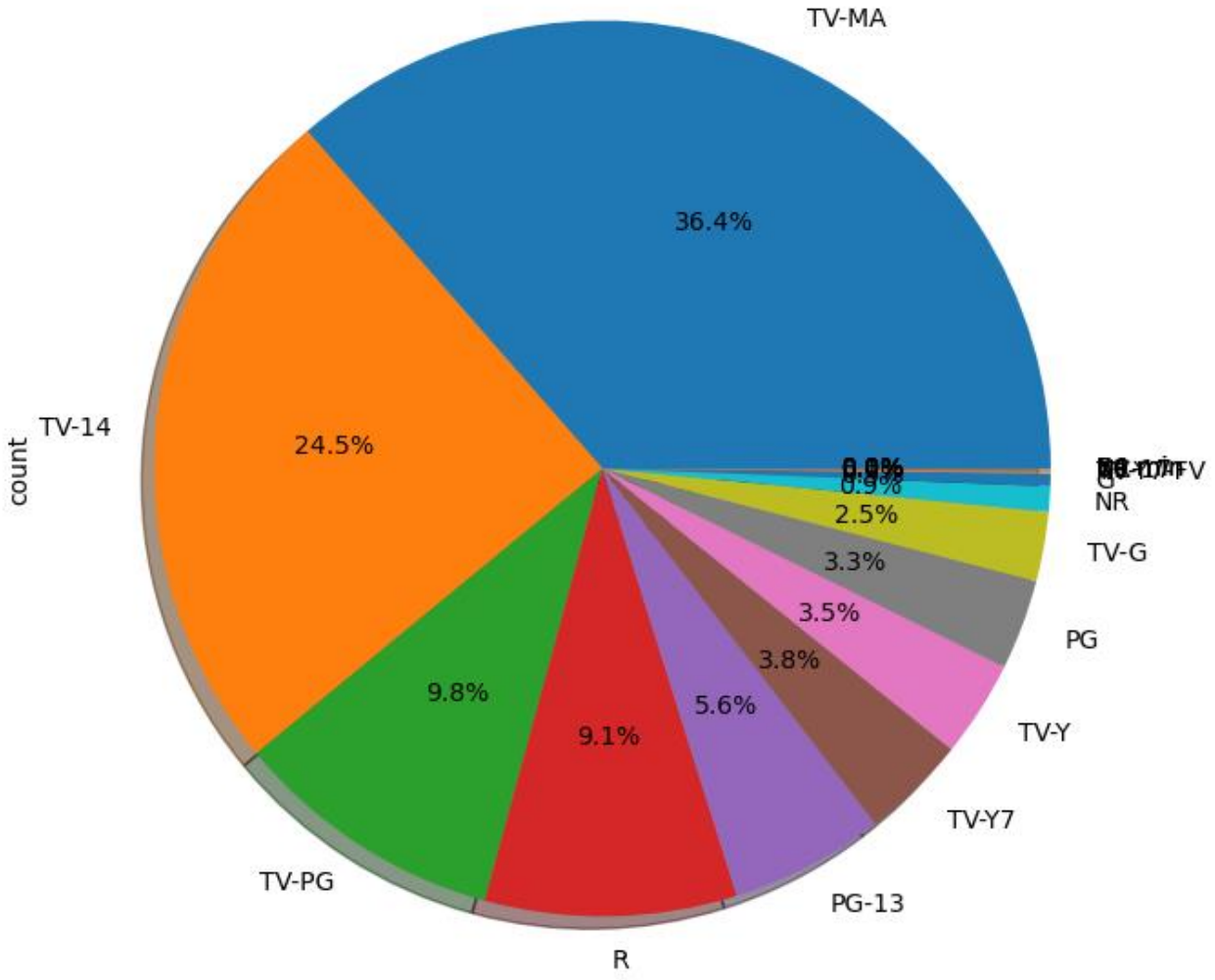

Fig 10 : The pie chart shows the distribution of ratings for Netflix content. The most common rating is TV-MA, accounting for 36.4% of the content, followed by TV-14 at 24.5%.

```
# Sample text or data (You can replace this with your actual text)
text = "Netflix has a variety of movies, TV shows, documentaries, and content for all kinds of audiences."

# Generate a word cloud image
wordcloud = WordCloud(width=1400, height=1400,
                      background_color='lavenderblush',
                      min_font_size=5).generate(text)

# Plot the WordCloud image
plt.figure(figsize=(8, 8), facecolor=None)
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")  # Hides the axis
```

```
plt.tight_layout(pad=0)

# Show the plot
plt.show()
```

Fig 11 : The word cloud highlights the key terms associated with Netflix, emphasizing its diverse content offerings. Words like "movies," "TV shows," "documentaries," and "variety" suggest a wide range of programming available on the platform. The prominent placement of "Netflix" itself underscores its central role in the context of the word cloud.

Country

In [39]:

```
plt.subplots(figsize=(25,15))
wordcloud = WordCloud(
                    background_color='white',
                    width=1920,
                    height=1080
                   ).generate(" ".join(df.country))
plt.imshow(wordcloud)
plt.axis('off')
plt.savefig('country.png')
plt.show()
```

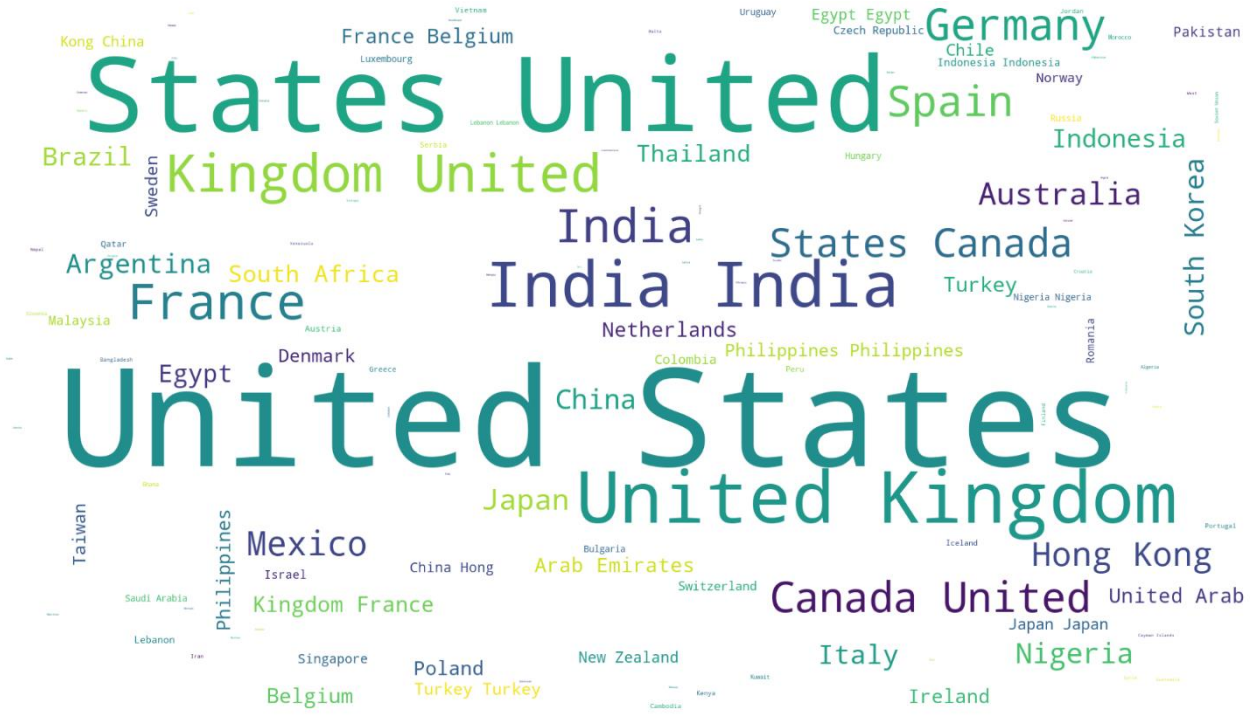

Fig 12 : The word cloud shows the top countries represented in the dataset. The United States, India, and the United Kingdom are the most frequently occurring countries.

Conclusion

This exploratory data analysis has revealed valuable insights into Netflix's content library and viewer preferences. Drama, Comedy, and Action emerged as the most popular genres, while Documentaries consistently received higher average ratings, reflecting their appeal to audiences. A sharp rise in content production post-2015 aligns with Netflix's global expansion, with the United States dominating contributions and countries like India and South Korea playing increasingly significant roles. Seasonal trends showed a peak in content releases during the fourth quarter, while longer-duration content, such as TV series, demonstrated higher viewer engagement. A positive correlation was observed between content ratings and viewer satisfaction, highlighting the importance of quality in driving engagement. However, the dataset lacks viewership metrics, which could provide a more accurate measure of content success. Future research could incorporate user reviews and watch statistics, as well as leverage machine learning to predict content success based on attributes like genre, region, and release year. These findings offer actionable insights for optimizing Netflix's content strategy, enhancing audience satisfaction, and maintaining competitiveness in the dynamic streaming landscape.

References

1. McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* O'Reilly Media.
   This book offers an introduction to using Python for data manipulation and analysis, with a focus on Pandas and NumPy.
2. Seaborn Documentation. Retrieved from https://seaborn.pydata.org/
   Official documentation for Seaborn, a Python library for creating advanced visualizations, ideal for EDA.
3. Python Crash Course for Data Science by Corey Schafer (YouTube series).
   YouTube tutorials explaining data analysis concepts using Python, covering Pandas, Seaborn, and Matplotlib for beginners.
4. DataCamp's *"Exploratory Data Analysis with Python"* Course. Retrieved from https://www.datacamp.com/
   An interactive course for students to learn data analysis techniques in Python, including data cleaning, transformation, and visualization.
5. Kaggle (Dataset Resource). Retrieved from https://www.kaggle.com/
   Kaggle offers open-source datasets, including the Netflix dataset, and community notebooks that demonstrate data cleaning and analysis techniques.
6. Pandas Documentation. Retrieved from https://pandas.pydata.org/
   Official documentation for Pandas, covering data handling, cleaning, and basic analysis tasks.
7. Matplotlib and Seaborn Tutorial on Real Python. Retrieved from https://realpython.com/
   Comprehensive tutorials on creating data visualizations with Python, useful for representing trends in Netflix data.
8. Statistics for Data Science Tutorial on Dataquest. Retrieved from https://www.dataquest.io/

This tutorial introduces basic statistical techniques for data analysis, covering concepts like central tendency and correlations.

9. "A Gentle Introduction to Exploratory Data Analysis" on Towards Data Science. Retrieved from https://towardsdatascience.com/
   A beginner-friendly guide to EDA, with a step-by-step approach to data inspection, cleaning, and visualization.
10. YouTube: *"Data Cleaning and Preprocessing with Python"* by freeCodeCamp.
    A hands-on tutorial that walks through data cleaning and preprocessing using Python, including handling missing data and outliers.
11. Jupyter Notebooks on Google Colab. Retrieved from https://colab.research.google.com/
    An accessible platform for practicing coding and data analysis, especially helpful for running EDA tasks in a browser environment.
12. Few, S. (2012). *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* Analytics Press.
    An introductory guide to creating effective graphs and tables, emphasizing clarity and insight in visualizations.
13. *"Introduction to Data Visualization with Python"* by DataCamp. Retrieved from https://www.datacamp.com/
    Covers different types of visualizations, ideal for students looking to understand and represent data patterns through graphs and charts.
14. Google, Kaggle, ChatGPT.
    Utilized as additional resources for gathering information, tools, and guidance during project development.

GITHUB REPOSITORY LINK:

https://github.com/TharunKumar0608/Netflix-Movies-and-TV-Shows